DOCUMENT RESUME

ED 457 196 TM 033 294

AUTHOR Giraud, Gerald; Impara, James C.
TITLE Making the Cut: A Qualitative Inquiry into the Setting of Cut Scores on School District Assessments.
PUB DATE 2000-04-00
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Elementary Secondary Education; Qualitative Research; School Districts; Scoring; Standards; *Teachers; Workshops
IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

This study investigated the cut score setting process as it occurred in two large Midwestern school districts, focusing on how the teachers who were the instruments by which cut scores were set experienced the process. Eight standard setting workshops using the Angoff approach were observed. Workshops for mathematics, reading, or writing at grades 2, 5, and 8 involved panels of from 24 to 28 teachers, and the ninth grade workshop involved 15 teachers. In addition to observation data, researchers held interviews with eight teachers who participated in the workshops and focus groups with five teachers after the grade-2 writing workshop and with three teachers after the grade-5 mathematics workshop. Some teachers answered questions after the fifth grade reading and writing workshops. In all of these workshops, teachers made judgments that resulted in cut score recommendations. These participants had volunteered for standard setting or had been recruited for their skills and cooperation. They came to the process with a willingness to help the district and the students, and they wanted to do well. These teachers were sensitive to the training provided in the Angoff workshops and were aware of district political and economic concerns. They were interested in setting cut scores that were in agreement with the district's goals for acceptable scores. Teachers relied on the definitions of the target examinee and the training provided in cut score workshops as clues to the policymakers' goals. The training these teachers received appeared to contribute to reliability. Several recommendations are made for improving the standard setting process. (Contains 16 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 457 196

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
G. Giraud
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Making the cut: A qualitative inquiry into the setting of cut scores on school district assessments

Gerald Giraud
Nebraska Methodist College

James C. Impara
University of Nebraska, Lincoln

Paper presented at the Midwest Educational Research Association 2000 Annual Meeting, Chicago, Illinois

TM033294

BEST COPY AVAILABLE

Studies of the Angoff method and its several variants have attended to the behavior of judges in workshops in terms of the judgments they make, and the resulting cut scores. Different types of training and marking strategies have been compared in terms of resulting cut scores or cut score ranges, and in terms of the dispersion of individual judge's cut scores (e.g. Plake & Impara, 1996; Plake & Giraud, 1997; Reid, 1991). The concepts of intra and inter judge reliability in item performance decisions have been considered, and suggestions made about how to maximize them (Plake, Melican, & Mills, 1991). Much of the inquiry into cut score setting has focused on issues of reliability (e.g. Norcini, et al. 1987; Reid,1991; Fehrman, Woehr and Arthur, 1991; Mills, Melican, and Ahluwalia, 1991; Kane, 1994; Plake, Impara, and Irwin, 1999). Reliability has been examined through quantitative analysis, with little attention to the underlying thinking of judges. No study has examined the subjective experience of judges as they are trained and provided with feedback data. No one has asked judges what they are thinking about as they make the judgments that are the basis of the cut scores that are derived from the process. Shepard (1994) summarizes the literature this way:

> [I]t can be concluded that two randomly equivalent panels of judges, led through the same procedure, will produce acceptably similar results when the judges' estimated standards are averaged. Such reliability studies tell us almost nothing about the substantive integrity of the resulting standards. (pg. 156) On the practical side, judges seem to have no difficulty following directions and implementing the Angoff procedure. (pg. 157)

The purpose of this study was to investigate the cut score setting process as it occurred in two large Midwestern school districts, and to do so with a focus on how the teachers who were the instruments by which cut scores were set experienced the process.

The response of judges to various aspects of the Angoff method can inform practitioners in the selection and refinement of training and feedback methods and content. Investigation of the process from the point of view of participants, along with a critical examination of the setting and the interactions of judges and presenters, can provide new insight and suggest points of focus for future investigation.

Research questions

Although past research on setting cut scores has focused on issues of reliability, little has been done to examine how judges in Angoff (1971) process experience the act of participating and making judgments. This study focuses on teachers who serve as judges in k-12 school district assessment cut score setting processes. The following questions guided the inquiry: What happens in Angoff cut score setting processes in school districts? How do teachers understand and respond to the key parts of the cut score setting process? How do they understand the task they are being asked to perform?

Method

The Workshops

Within the parameters of this study eight standard setting workshops were observed. These workshops were conducted for the purpose of suggesting cut scores on examinations that would classify students as needing instructional assistance beyond classroom instruction in specific domains (e.g. reading, mathematics) at various grade levels: 5th grade mathematics, reading, and writing; 2nd grade reading, writing and mathematics, 8th grade mathematics, and 9th grade mathematics. The workshops for grades 2, 5, and 8 were held in one district, and the 9th grade workshop in another. Workshops for grades 2, 5, and 8 involved panels of from 24 to 28 teachers, while the 9th grade workshop involved 15 teachers.

The teachers who served as judges in the workshops were experienced teachers. All had tenure (at least three years of experience) and were currently teaching the grade level and subject matter of the test for which a cut score was being determined.

The workshops were not random samples of the many workshops that are conducted for purposes of cut-score setting: They are examples of workshops conducted or overseen by persons whose methods are grounded in the theory and practice of cut score setting, particularly the Angoff method.

Data collection

Observation data. The workshops were observed, and notes recording the observations were made, both on a laptop computer and in written form. Participants were observed as they arrived, during formal activities of the workshops, during breaks and lunch periods and as they left the workshops. Observations attended to the actions of presenters and panelists, and to what they said. The principal investigator also talked with other persons who were present, such as school district personnel who were observing the workshops for their own purposes, asked questions about what was observed, or engaged them in discussion about what was occurring in the workshops. Questions asked and discussions were about a) particular teachers, b) the political climate of the district, c) the students of the district, or d) the actions of participants or what participants said.

The note-recorded observations of the workshops were the primary observation data sources for this study, but meetings between consultants and school district personnel that were for the purpose of planning the workshops were also observed. These observations served as background for the focus of the study, and were used to establish context for observations and analysis.

Interview data. Eight teachers who participated in workshops were individually interviewed: three 9th grade mathematics teachers following the 9th grade workshop, one 5th grade teacher following the 5th grade reading workshop, one 5th grade teacher following the 5th grade writing workshop, two 2nd grade teachers following the 2nd grade reading workshop, and one 2nd grade teacher following the 2nd grade writing workshop. Focus groups were conducted following the grade 2 writing (five teachers) and grade 5 mathematics (three teachers) workshops.

All interviews, both focus group and individual, were recorded and later transcribed. Finally, some teachers also responded in writing to questions following the 5th grade reading and writing workshops.

Interview protocol

Questions asked during the interview followed a general theme of open-ended queries intended to illicit teacher reactions to the workshop. The interviews began with this question:

> Think about when you were asked to participate in the workshop. What did you think? What were your feelings about being asked?

The interviews proceeded in like manner through the events of the cut score process, and as dictated by the teachers being interviewed, with the interviewer following up on topics raised by the teachers being interviewed. Interviews lasted approximately one-half hour. The interviews were recorded, and the recordings transcribed.

Member Checks. Three teachers who were interviewed were sent transcripts of their interviews, along with write-ups of preliminary observational descriptions and analysis. No substantive suggestions for change or clarification came from these member checks.

Analysis of data

The data collected from observations and interviews were analyzed by first compiling them into descriptions of what occurred in the cut score setting processes. After describing what occurred, a distillation of salient issues was developed.

Interview and focus group data analysis

Interview and focus group data were analyzed through a process similar to that suggested by Moustakis (1994). First, interview and focus group transcripts were read through entirely to get an overview of the teachers' reactions to the process. Next, a line by line reading was done and salient statements were extracted, recorded and numbered. Three hundred and seventy nine statements were extracted. Each statement identified as salient was then evaluated on two standards: 1) Did the statement aid understanding of how teachers reacted to, understood, or experienced the standard setting process, and 2) Was it possible to abstract and label the statements into conceptual categories?

The statements derived from the two step evaluation were clustered into theme categories and evaluated for redundancy, frequency, and emphasis in the context of the complete record, including interview transcripts and the totality of observations, understanding and knowledge of the standard setting contexts that were the focus of this inquiry. Statements that were meaningful in the context of the complete record were again examined and placed into a thematic framework. A description of teachers' experience and reactions to the standard setting process was constructed around this framework.

Results

A description of the cut score setting process

The literature of the Angoff (1971) method suggests that judges (teachers, in this instance) be 'trained' to understand the target examinee (e.g. Berk, 1986, 1996; Reid, 1991). In the processes studied here, teachers were trained on the Barely Master concept, where the Barely Master student was the target examinee.

Teachers were provided with definitions of levels of mastery of the domain for which the cut score was being set. These definitions took the following form (modified according to the grade level and domain of interest):

> Nonmaster- The student needs substantial assistance to accomplish appropriate tasks. The student probably needs special interventions to succeed. The student still is in the process of learning skills and strategies that are required in most applications.
>
> Barely master- The student can complete some appropriate tasks independently and can get by on other tasks with normal help from the teacher or other adult. This student is one who can do most assigned tasks, but with some difficulty given some material.
>
> Master-The student can accomplish many appropriate tasks with minimal assistance from the teacher and can perform most tasks encountered in daily experience.
>
> Definite master-The student is accomplished and can comprehend and appreciate a variety of material encountered in daily experience.

These definitions were developed by the workshop facilitators, in collaboration with school district staff, to assist teachers in conceptualizing the Barely Master Student (BMS), who was the target examinee. Consideration was given to the desires of policy makers in deciding which level was the target examinee, and also in deciding the exact definition of the target level of mastery. The school districts in the current study wanted to identify as non masters those

students who needed help beyond what a teacher could give in the classroom to 'succeed' in the domain of interest. Students who obtained a score higher than the cut score were to be considered at least minimally masters.

During the Angoff workshops, these definitions were elaborated through discussion and training. One device that the facilitators used to illustrate the task at hand was to display a number line, one end of which was the maximum score, opposed on the other by the lowest possible score. The facilitator then put mastery on the template of the number line, saying: "Students who score here [indicating the maximum score] are probably definite masters, and here non masters [indicating a point near the lowest score]. The master students will likely score in some middle range, and here is the borderline between mastery and non mastery [usually indicating a point about one-fourth of the length of the line away from the lowest score]." If this demonstration was based on a number line of from 0 to 100, the indicated borderline would be at about 25 points. This demonstration was not intended to tell teachers the cut score, but to help them conceptualize the task at hand.

Following these definitions, a discussion about the barely master students' skills relative to the domain of interest was facilitated (see Mills, et al. 1991). Lists of tasks that would be hard and easy for the Barely Master Student were elicited from the teachers, who had been asked to put themselves 'into the skin' of a particular Barely Master Student they have in their classroom.

Teachers then examined each item on the test for which a cut score was desired, and made a performance estimate for each item. Specifically, the teachers judged whether a target examinee (a BMS) would correctly or incorrectly answer each multiple choice (or other dichotomously scored) item. For constructed response items, the teachers selected examples of work from throughout the score range that reflected the performance of the BMS. The item judgments were summed to determine individual teacher's cut scores, and these individual cut scores were averaged to determine a 'final' cut score. After making judgments on all test items, teachers were provided information about how all examinees (not just the barely masters) performed on each item, and about the impact of the cut score derived from this first set of judgments (i.e., how many examinees would be classified as non masters given this cut score). Finally, teachers re-examined the test items and made judgments of performance upon which the final cut score recommendation resulting from the Angoff method was made. In the cut score setting processes that were the focus of this study, policy makers were given a range of cut scores from which to choose a final operational cut score.

Description of teacher experience(Based on interviews and focus groups)

Reason for Participating

Teachers agreed to participate in the standard setting process because they felt knowledgeable and competent, and because they viewed it as an opportunity for professional development. They wanted to make a difference, to be part of the assessment decision making process in the district.

Reaction to the process

Affect. Teachers who participated had a positive affect toward the workshop.

Importance. Teachers believed that they were involved in an important process, one that was not to be taken lightly and involved high stakes for students, teachers and the district.

Affirmation of teachers. Teachers felt that being asked by the district to participate was affirming of their competence and knowledge. They felt they were treated well by the workshop organizers, and this made them feel valued.

Presenters. Teachers found the workshop presenters to be open, honest, and knowledgeable, although some felt a barrier between outside consultants and themselves. This was quickly resolved, however, as the workshop progressed.

Teachers who attended. Teachers who participated felt they were part of a special group of teachers, who 'had a clue', who were interested in the assessment process, had strong opinions, and wanted to make an impact on decisions. They were likely to be involved in school and district activities outside of their classrooms.

Some teachers were uneasy about the competency or motives of other teachers on the panel, pointing out that they (other teachers) did not understand what was expected in the workshop, or were there to make an appearance of upholding the standards.

Training and Interaction among teachers. Teachers felt the workshop training helped them to focus on and understand the task. Teachers liked the discussion among themselves about the Barely Master Student, and many felt it was the most important part of the workshop. They took into account what other teachers said in these discussions when forming their own understanding. Teachers would have liked more interaction among themselves.

Compliance. Expectations were made clear for most teachers. They were committed to following instructions and doing their part in setting the cut score. They wanted to do a good job for the district.

The Barely Master Student (the target examinee)

Conceptualization and Workshop Definition. The teachers conceptualized the Barely Master Student by attending to the definition that was provided by the workshop facilitators, even when they had doubts about the appropriateness of the definition. Some felt the concept of barely master was unclear, and struggled throughout the workshop to understand and apply it. The facilitators suggested a device for conceptualizing the Barely Master Student: Identify a Barely Master Student whom you know and keep that student in mind as judgments are made. Teachers accepted this method, and thought of one or more students in their classes who they believed were barely master, according to the definition provided by the facilitators.

Characteristics of the Individual target examinee. Teachers described the characteristics of the Barely Master Student. This was a student, according to the teachers, who was in a range of ability, between almost qualified for special education and 'on the page (of mastery) but not solid'. These students need some special attention, a personal relationship that motivates them to perform. They want and need praise.

Teachers describe these students as tending to underestimate their own abilities, often frustrated, used to failing or barely succeeding, and reluctant to try too hard. They are fragile, from low income and unstable homes. Some are nervous and worried about school expectations.

These students need help, according to teachers. Most of these students cannot achieve mastery without extra help, but some can do it more often than not on their own. The help required was estimated as between 'some one sits there and works with them', and 'going to need a lot of extra assistance' to 'will be okay with the help I can give in the classroom'.

Making Judgments

Thinking about the Barely Master Student. As teachers made judgments about the performance of the Barely Master Student on test items, they kept in mind one or several individual students whom they felt fit the definition of Barely Master provided in the training. They felt they could not be too optimistic about the performance of these students, and were thoughtful about whether they were accurately considering the performance of these students. Sometimes, the teachers found it difficult to set aside their expectations as teachers, and to instead focus on what the Barely Master would do on the test items.

Preconception of cut score. As teachers made their judgments about the performance of the Barely Master Student on test items, they sometimes considered what they believed to be a passing score, based on past experience. They thought of scores that they considered passing in their classroom tests, or on state assessments. These scores were usually higher than the conceptualized Barely Master Student would obtain. Teachers identified 70 percent correct as

passing, or else a score of 3 on a 5 point scale. Even though they were experienced with tests that had a preset passing score (like 70 percent correct or a 3 on a 5 point scale) most teachers were willing to put that aside as they made judgments.

Taking into account error. Teachers took into account the possibility that students would not perform as well as they could on the test items. This was reported as a comforting factor when scores were lower than teachers would have liked, given their preconceived notions.

Expertise. Teachers' knowledge of scoring the assessment (especially the writing assessment) was sometimes a confounding factor in their attempt to follow the directions for making judgments. They sometimes wanted to assign score values to written passages, for example, and then pick a score they thought was a reasonable passing score. However, most teachers felt that their prior knowledge of the assessment gave them confidence in making judgments.

The influence of feedback

Understanding. Teachers were skeptical about how other teachers understood the cumulative score data and item difficulty information given between rounds. Comments included "I think maybe 25 percent of teachers understood it", "I had trouble understanding why they were doing it", and "the feedback looked confusing when I first saw it". However, "as time came on, they understood it more".

Confirmatory use and influence on judgments. Although they expressed some concerns about undstanding the feedback, teachers used the item difficulty information as confirmatory data. If their judgments on item performance of the BMS were in line with the feedback information, they felt that their judgments were confirmed. They felt good when this happened. When their judgments were not congruent with the performance of students on some items, they would reread and reconsider those items, sometimes changing their initial judgment. This was especially true if they were unsure of their initial judgment before seeing the feedback data. If a teacher felt strongly that the Barely Master Student they had in mind would fail the item, then they might not change their judgment, even if there was wide disagreement between their judgment and operational item performance.

Teachers reported that they were less influenced by the impact data provided between rounds than by the item difficulty information. Although some teachers said they were interested in it, only a few reported changing their judgments in an attempt to influence the cut score. This occurred when a teacher felt the cut score was too low relative to their perception of an appropriate failing rate (when a higher failing rate was expected), or to their perception of an appropriate passing score.

Concerns about the impact of the cut score

For students. Even though they often did not change judgments based on the impact data, teachers were concerned about the impact of the cut score that would result from their efforts. They thought about what would be done to remediate the students who failed to obtain the cut score. They were also concerned about whether the cut score would be too high, resulting in an unrealistic expectation for student performance, or more commonly too low, resulting in students who needed help not getting it. A particular worry was that the Barely Master Student, the one who would just pass, would not receive needed assistance.

For teachers. The impact of the cut score on teachers was also a concern. The decision to set passing scores was seen as controversial among teachers in general because of "rumors [that the cut score and student performance] is going to determine the amount of money teacher's make".

For the district. The teachers recognized that the district (school board) had paid a lot of money for consultants to facilitate the workshop, and also that the process was a political one.

They were aware that the district had to make decisions based on economic and political concerns, and might raise or lower a recommended cut score accordingly.

Discussion

In the school district cut score setting processes studied here, teachers made judgments that resulted in cut score recommendations to policy makers . If Shepard (1979) is correct, then the validity of the cut scores derived from these processes inheres in the wisdom of the teachers. This study examined how teachers experienced the cut score setting process, how they made judgments about the performance of the minimally competent student on test items, and how they were influenced by the activities and context of the cut score setting processes in which they participated.

The particular teachers who participated in the cut score setting processes studied here were not a random or even representative sample of all teachers in their districts, but were teachers who either volunteered because they had an interest in participating, or were recruited based on their record of cooperation and participation in district processes. The teachers described themselves as knowledgeable and involved in their schools and the district, and said the 'other' teachers not in attendance were less knowledgeable and involved than they.

Teachers came to the process with a willingness to help both the district and the students of the district. They wanted to understand what was expected of them and to do their assigned tasks as they understood them. They felt affirmed by being asked to participate in what they saw as an important process. Even when they perceived that the process was leading to a result that they questioned or even strongly disagreed with, they set aside their own views of competence and appropriate standards and conformed to what they perceived to be the expectations of the district, as communicated by the process facilitators.

The processes were directed by psychometricians, who were introduced to the teachers as experts in cut score setting methods. Thus, the teachers were immediately confronted with an authority who had specialized knowledge that the teachers did not have. These specialists were to direct them in the task they were to undertake. In their classrooms, teachers are the authorities, and they arbitrate domains of knowledge. Teachers are accustomed to knowing the right answer, and judging the rightness or wrongness of answers that students provide. In the workshop, psychometricians stood in the front of the room, and arbitrated the cut score setting process. Teachers were situated in relation to the facilitators as students are situated in relation to the teachers in the teachers' classrooms. The teachers who served as judges in these workshops were therefore in the role of being directed by experts who knew the right answer to the question of setting cut scores. Teachers were told when they could talk to each other and when they could not. Teachers said in interviews that interaction with other teachers about the concept of barely master was the most useful part of the process, however after this discussion they were not free to interact as they made operatonal judgments.

Teachers were sensitive to the training provided in the Angoff workshops studied here. The training involved first defining the target examinee, then operationalizing the performance of the target examinee on an assessment through a discussion among teachers directed by the workshop facilitators. Teachers reported that they often thought of what other teachers had said during this discussion as they made judgments later. After training, teachers were asked to think of a particular student whom they knew who fit the description of the target examinee that had been constructed in the training, and to keep this student in mind as they made judgments about performance on test items. They were asked to think about what this student would do on test day, and not to consider what a minimally competent (Barely Master) student should or even what the student they had in mind could do. Teachers indicated that they did think of one, or

several, students whom they knew who fit the definition and operationalization of the training. They acknowledged thinking about what the student they had in mind would do on test day, not what the student should or even could do. Teachers reported relying on the definition provided to guide them in their judgmental thinking.

The teachers who participated were aware of district political and economic concerns. They knew that the school board, who would make the final decision about the cut score, would be concerned with public perception, as well as with the resources required to remediate the students categorized as not competent by the cut score. Sometimes these concerns were at odds in terms of a 'desirable' cut score. For example, a cut score that was too high might fail too many students, making the district look bad and demanding too many resources for remediation. On the other had, a score that was too low might be viewed as not rigorous enough, and therefore be politically untenable. In addition to a nonspecific concern about how a particular cut score resulting from their judgments might be perceived by policy makers, teachers were also aware of more specific expectations. One of the districts had a policy, for example, that defined a passing performance as 70% correct, set by the school board of the district. Teachers were concerned that the cut score that resulted from their efforts might not be acceptable to policy makers, that they might look 'stupid' if their judgment did not conform to policy maker's expectations. In one district, teachers had been told on previous occasions that a score of 3 on a 5 point scale was an acceptable score on a state mandated writing test. They wondered how the cut score that resulted from the Angoff (1971) workshop would be reconciled with this acceptable score. Teachers were resigned to letting the school board figure this out, but they still reported thinking that the training was pointed at some district desired cut score or percentage of not competent students. When in doubt, they turned to the definition of the target student that was supplied by the facilitators, in the belief that it reflected the needs and expectations of the district.

Teachers were also provided with feedback data in two forms: item performance data and cut score impact data. Teachers used the item performance data as a confirmatory device. If the data disconfirmed their initial judgments by being incongruent with their item performance judgment, teachers reexamined the items in question and considered changing their judgments. Sometimes they did change their judgments. Most teachers did not report being influenced by the impact data, although a few did say they changed some item judgments in an effort to raise the cut score.

In some of the workshops, teachers seemed not to attend to the feedback at all, and simply recorded their initial estimates as second round judgments. This occurred in the latter stages of workshops where interim cut scores and attendant feedback data were reported for portions of the assessment of interest. It could be that teachers were satisfied by data presented earlier in the workshop that their judgments were satisfactory.

Assertions

Teachers bring to the standard setting process knowledge of political and economic issues, as well as preconceived notions about passing scores. Political issues involve the expectations of the larger community in terms of student performance and outcome measures, such as fail rates on important assessments. Economic issues involve the resources that districts have available to provide remediation for students categorized as not competent by the cut score. As teachers make judgments about the performance of the target examinee, this prior knowledge is either considered or set aside. In the cut score setting processes studied here, teachers tried to discern what it was that the district wanted in terms of a cut score. Teachers used the knowledge they had about the political and economic situation in the district as they attempted to align their thinking with the cut score setting process. They relied on the definitions of the target examinee and the training provided in the cut score workshop as clues to the policy makers' goals.

These assertions fit with findings related to reliability suggested by the research into the behavior of judges as they participate in formalized processes such as the Angoff (1971) method. The teachers involved in the processes studied here did attend to the training, and did follow the directions of the facilitators as they (the teachers) understood them. If they came to the process with some private notion of an acceptable cut score, they could put that aside and adopt what they believed to be the view of the school district, as communicated in the training. Teachers valued and took into account the statements of other teachers as they made judgments. Thus, training had the desired effect of encouraging the judges to think about the judgments made in the process in a standard way. The training and formalization of the process was powerful enough even to overcome strongly held views about the definition of competence and desirable cut scores. This likely results in the reliability that psychometricians value.

However, the power of training and discussion among judges to alter judges behavior calls to mind the issues raised by Fitzpatrick's (1989) review of the social science literature of group interactions. Fitzpatrick called attention to the possibility that some judges could unduly influence the decisions of others in the process. Teachers reported, for example, that the comments of other teachers were an important component in their judgment making. Whether their resulting judgments were more extreme than they might have been, a danger suggested by Fitzpatrick's review, is undiscovered in this study, but evidence for the affect of others' opinions was found. Further, the teachers reported relying on the written definition of just competent (barely master) provided by the workshop facilitators as they made judgments. Whether teachers were providing judgments based on their own expertise, or conforming to their understanding of views of other teachers or the desires of the district policy makers is the difficult question that arises from this inquiry. In our view, teachers who were interviewed attempted to conform to the expectations, as they understood them, of the district policy makers as communicated through the workshop training.

Recommendations for practice of standard setting

In standard setting processes in school districts, the definition of the target examinee used in the training process must be carefully constructed by the policy makers who desire the cut score. Because there is no standard definition of just competent, and because teachers seem to rely on the definition given by the facilitator in making judgments, it is important that a definition is provided that satisfies the concerns of the district in which the cut score is to be used. A definition that is not carefully considered is likely to yield a result that is unintended.

A carefully considered definition is important even in the context of allowing teachers to discuss and formulate a behavioral description of some level of competence in some domain of interest, because even a global definition, as suggested by Mills, et al. (1991) defines the focus of the discussion. The definition of the target examinee used in the processes studied here were a bit more descriptive than global, and included clues related to how much assistance a just competent (barely master) student would need to perform tasks in the domain of interest successfully. Teachers relied on this definition as they made judgments, a finding that points to the influence of the definition provided by the facilitators.

A second recommendation is that when the results of the standard setting workshop are presented to the policy makers who will set the standard, it would be prudent to include in the presentation some description of what the experts (teachers) thought about as they made the judgment upon which the suggested cut score is based. It would be particularly helpful to include a description of how the teachers defined the barely master, or just competent student who was the target examinee. Such a description would allow policy makers the opportunity to determine whether teachers had in mind the level of performance that fit the policy makers' notion of just

competent. The presentation would include the definition used to initiate the discussion along with illustrations of the characteristics defined by the teachers in their discussion.

A final and general recommendation is that standard setting practitioners attend to the larger context of setting cut scores in school districts. Teachers who serve as expert judges on Angoff panels bring with them a contextual understanding of the political and economic realities of the district, a relationship with district authority, and a notion of what constitutes competent student performance. Policy makers have an interest in the cut score outcome in terms of political ramifications and economic impact. These factors will likely affect the meaning of a cut score that is derived from the judgments of teachers and policy makers. Consultants who facilitate the cut score setting process should be careful to take these issues into account as they plan and direct the process that results in a recommended cut score, if only in terms of encouraging policy makers to carefully consider the implications of setting a cut score on an assessment for purposes of taking action.

References

Angoff, W. (1971) Scales, Norms and Equivalent Scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed., pp.508-600) Washington, DC, American Council of Education.

Berk, R. (1986) A consumers guide to setting standards on criterion referenced tests. Review of educational research, 56, 137-172.

Berk, R. (1996) Standard setting, the next generation. Applied measurement in education, 9, 215-235.

Fehrmann, M., Woehr, D. and Arthur, W. (1991) The Angoff cutoff score method: The impact of frame-of -reference rater training. Educational and psychological measurement, 51, 857-872.

Fitzpatrick, A. (1989) Social influences in standard setting: The affects of social interaction on group judgments. Review of educational research, 59, (3) 315-328.

Kane, M. (1994) Validating the performance standards associated with passing scores. Review of educational research, 64, 425-461.

Mills, C. , Melican, G. and Ahluwalia, N. (1991) Defining minimal competence. Educational measurement: Issues and practice, 10, 7-10.

Moustakas, C. (1994) Phenomenological research methods Thousand Oaks, CA: Sage pp. 120-153 .

Norcini, J., Lipner, R., Langdon, L. and Strecker, C. (1987) A comparison of three variations on a standard-setting method. Journal of educational measurement, 24, 56-64.

Plake, B., Giraud, G. (1998). Effect of a modified Angoff strategy for obtaining item performance estimates in a standard setting study. Paper presented at the 1997 Midwest Educational Research Conference, Chicago.

Plake, Barbara S.; Impara, James C.; Irwin, Patrick. (1999) Validation of Angoff-based Predictions of Item Performance. . Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

Plake, Barbara S.; Impara. James C. (1996). Intrajudge Consistency Using the Angoff Standard-Setting Method. Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).

Plake, B., Melican, G., and Mills, C. (1991) Factors influencing intrajudge consistency during standard setting. Educational measurement: Issues and practice, 10, (2) 15-25.

Reid, J. (1991) Training judges to generate standard-setting data. Educational measurement: Issues and Practice, 10, 11-14.

Shepard, L. (1979) Setting Standards. In Bunda, M. A & Sanders, J.R. (ed.) Practice and Problems in competency-based measurement, National Council on Education Monograph.

Shepard, L. (1994) Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress achievement levels. Paper published in the Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments. Washington, D.C.